International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 1, Feb 2014, 133-140
© TJPRC Pvt. Ltd.

INTELLIGENT RECOMMENDATION SYSTEM USING CLUSTERING AND CLOSED SEQUENTIAL PATTERN MINING

MITCHELL R. D'SILVA[1] & DEEPALI VORA[2]

[1]M.E. Student, Information Technology, Vidyalankar Institute of Technology, Maharashtra, India

[2]Information Technology, Vidyalankar Institute of Technology, Maharashtra, India

ABSTRACT

The World Wide Web is a massive source of information that is gaining wide usage due to constant availability of huge subject content, ability to share resources, better presentation, dynamically changing content, flexibility of time, discussion facility through chats, blogs and forums etc. However, many web based learning systems lack assessment of user activities and learners are confused by huge number of web pages in the website. Searching and visiting several web pages to get the relevant content also increases the network traffic as well as the load on the server. The proposed system aims to overcome the drawback of accessing relevant web pages from the huge number of similar web pages that are available in the website by providing automatic online navigation recommendations to the users. It uses techniques such as Pre-Processing, Rough Set Clustering, Sequential Pattern Mining using Prefix Span and Closed Sequential Pattern Mining using post pruning strategy. The recommendations generated are evaluated for accuracy based on four parameters namely Precision, Recall, Miss Rate and Fallout.

KEYWORDS: Preprocessing, Rough Set Clustering, Upper Approximation, Sequential Pattern Mining, Prefix Span, Closed Sequential Pattern Mining

INTRODUCTION

The web has become an important information resource for people to search information and communicate across the corners of the world. However, users find it difficult to find relevant materials of their interest from this vast enriched source. Users visit various links in the website that they think are relevant until they find the desired information in one or more pages. This increases the browsing time as well as the network traffic. These problems can be solved by providing useful navigation recommendations to users based on previous user's browsing patterns. Using various web usage mining techniques we can predict the next page to the user. This makes it easy for the user to browse the website and find relevant content quickly. It also reduces the network latency by pre-fetching the recommended web pages [1].

The proposed intelligent navigation recommendation system is implemented using Rough Set Clustering using upper approximation, Sequential Pattern Mining using Prefix Span and Closed Sequential Pattern Mining using post-pruning strategy. These techniques are applied on the Web Server log data. The web log data is first pre-processed to filter out unwanted and ambiguous data and organize the relevant data before applying any of the data mining techniques. Rough Set Clustering is then used to form clusters of web transactions that have similar browsing patterns. The rough clusters are then upper approximated until it results in mutually disjoint equivalence classes. Sequential pattern mining technique called Prefix Span is then applied on individual clusters [1]. Prefix Span determines all frequent sequential patterns from the clusters of transactions. Sequential patterns whose support is greater than the specified minimum support are selected and then projected database of those sequences are created thereby providing the final output until no sequence with support greater than the specified value are available. Prefix Span generates a large number of patterns exponentially which becomes a problem when the database consists of long frequent sequential patterns. Closed sequential pattern mining is then used to reduce the number of patterns generated by Prefix Span by using the post pruning strategy. It eliminates all the patterns in the set whose supersets are present in the set and that have the same support as that of the superset [1]. The closed sequential patterns thus generated are stored in the database. Whenever a user visits “minlen” number of web pages, the user’s current web access sequencesare matched with the patterns in the database and all the matching patterns are provided as recommendations to the user. User can select and visit any of the web pages from amongst the recommended list of pages as per his/her interest.

Further, the proposed system evaluates the accuracy of navigation recommendations provided by the system based on four parameters namely Precision, Recall, Miss Rate and Fallout.

PROPOSED RECOMMENDATION SYSTEM DESIGN

Figure 1 shows the design of the proposed navigation recommendation system followed by the detailed explanation of each block in the design.

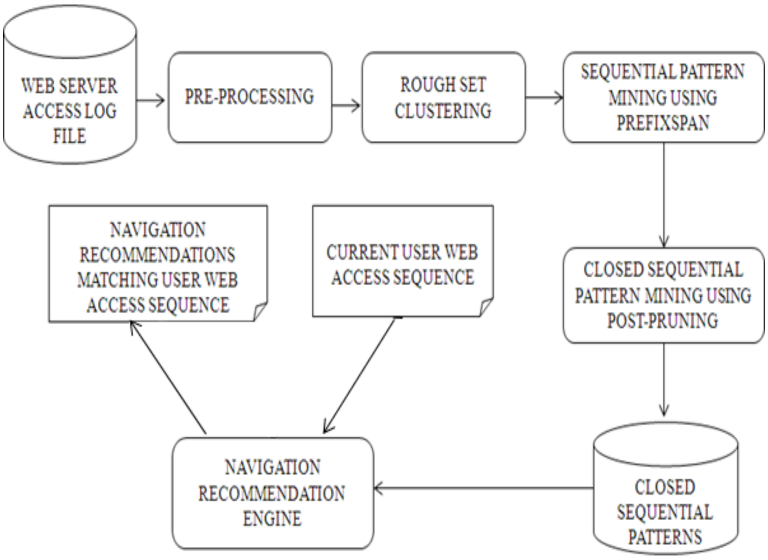
WEB SERVER ACCESS LOG FILE
PRE-PROCESSING
ROUGH SET CLUSTERING
SEQUENTIAL PATTERN MINING USING PREFIXSPAN
CLOSED SEQUENTIAL PATTERN MINING USING POST-PRUNING
CLOSED SEQUENTIAL PATTERNS
NAVIGATION RECOMMENDATION ENGINE
CURRENT USER WEB ACCESS SEQUENCE
NAVIGATION RECOMMENDATIONS MATCHING USER WEB ACCESS SEQUENCE

Figure 1: Proposed Navigation Recommendation System

The proposed navigation recommendation system operates in two phases namely offline phase and online phase which are explained as follows:

Offline Phase

Offline phase does not directly interact with the online user. The first six blocks operate in the offline phase. The previous users’ browsing patterns are recorded and stored in the database also called as the Web Server access log. The blocks in the offline phase operate on this log data to discover useful patterns from it which are then stored in the database to be used by the online phase.

Online Phase

Online phase directly interacts with the online users. The last three blocks operate in the online phase. In this phase, the current browsing sequence of each active user is sent to the recommendation engine, which then matches that access sequence with the useful patterns in the database and provides navigation recommendations to the users.

The various stages involved in the proposed navigation recommendation system areexplained in detail as follows.

Web Server Access Log File

A website is developed and hosted on http://www.techie-world.tk. This website named "Tech World" presents information about various Web Technologies like HTML, CSS, JavaScript, JSP, ASP and XML including quizzes for each technology. It comprises of 450 web pages and videos. The input data used for the proposed recommendation system is the web server access log of the above mentioned website. The data acquisition process takes place in 3 steps namely registration, login and recording entries while browsing the website.Each user is assigned a unique session id and a timestamp while logging into the website inorder to differentiate among the users. The web server records each user's details in the following 10 fields' namely remote, method, URL, agent, protocol, referrer, code, username, session id and timestamp.

Pre-Processing

The input to this block is the Web Server Access log file and the output is a set of pre-processed data arranged in a suitable format required for further analysis. Preprocessing is used to extract useful data from raw log data and arrange it in a form suitable for pattern discovery. Preprocessing includes two stages namely data cleansingand data structuration. Data cleansingconsists of removing useless requests related to irrelevant images, multimedia files, failed HTTP status codes and entries generated by robots. Data structuration consists of two steps namely user identification and sessionidentification. User identification identifies unique users from the log file based on IP address, browsers, operating systems or referrer page information. Session identification is used to divide the page accesses of each user into individual sessions using the method of timeout which is usually30 minutes [2][3].

Rough Set Clustering using Upper Approximation

The input to this block is the pre-processed web server access log data and the output is a set of clusters that have similar transactions.The proposed system uses rough agglomerative approach to cluster web user transactions because robust clustering methods are needed when user's browsing patterns are hidden in data with noise and outlier components.Web data clustering is the process of grouping web data into "clusters" based on similarity threshold value so that similar pages are in the same class and dissimilar pages are in different classes. In the proposed system, the similarity threshold value is set to 0.4 [1]. Rough Set Clustering helps to aid decision making in the presence of uncertainty. It classifies imprecise, uncertain or incomplete information expressed in terms of data acquired from experience.A rough set is defined by pair of sets which gives lower approximation and upper approximation of the original set [4]. The steps for Rough Set Clustering are as follows [1]:

- Construct a similarity matrixusing Jaccord's coefficient.
- Calculate the similarity class for each transaction where the similarity threshold value is taken as 0.4.
- Compute upper approximation. This step is repeated until the result of two successive iterations isthe same.

Sequential Pattern Mining Using PrefixSpan

The input to this block is the set of clusters generated by Rough Set Clustering technique and the output is a set of frequent sequential patterns. Prefix Span is a pattern-growth algorithm used for mining complete set of sequential web access patterns from the sequential database. Its main goal is to project frequent prefixes in the sequence rather than projecting the sequence database by considering all possible occurrences of frequent subsequences based on predefined minimum support count [5].

In the proposed system, each cluster is considered as a sequence database. The several transactions in each cluster are considered as elements and the pages belonging to each of these transactions are considered as the items of those elements. The minimum support count is set as 1% of the total number of transactions. Each cluster generated by Rough Set Clustering technique is passed individually to this block. Prefix Span is then applied cumulatively on all transactions within a single cluster to generate a set of frequent sequential patterns. The steps for Prefix Span are as follows [1]:

- Find all length-1 sequential patterns from each transaction.
- Construct projected database for each length-1 sequential pattern.
- For each projected database, find the subsets of sequential patterns. This step is repeated until no more frequent subsequences can be generated from that particular projected database.

Closed Sequential Pattern Mining Using Post-Pruning Strategy

The input to this block is the set of frequent sequential patterns generated by Prefix Span technique and the output is a set of closed sequential patterns. Closed sequential pattern mining is used to mine closed sequential web access patterns from the complete set of sequential web access patterns (FS) using post-pruning approach.Closed frequent sequential pattern is defined as follows [5]:

$CS = \{\alpha \mid \alpha \in FS \text{ and } \beta \in FS \text{ such that } \alpha \subseteq \beta \text{ and support } (\alpha) = \text{support } (\beta)\}$

Each sequential web access pattern Fs_i, in the pattern set FS is compared with the other patterns in the set for instance Fs_j. The pattern Fs_i is removed from the pattern set FS, if and only ifthe support of both web access patterns Fs_i and Fs_j are the same and Fs_i is a subset of Fs_j [5].

Closed Sequential Pattern Database

This block is used to store the closed sequential patterns generated by performing closed sequential pattern mining using the post pruning strategy on the frequent patterns generated by Prefix Span. This block acts as an input to the recommendation engine. The patterns in this database are used by the recommendation engine to provide navigation recommendations to the users thereby saving users browsing time as well as reducing the network traffic and bandwidth requirement.

Current User Web Access Sequence

The users register themselves on the website and then login to the website. This block represents a user's current browsing sequence. This browsing sequence is passed as input to the navigation recommendation engine.

Navigation Recommendation Engine

The input to this block is the current user's web access sequence as well as the closed sequential pattern database and the output is a set of navigation recommendations to the user. It uses the patterns in the closed sequential pattern database for matching and generating navigation recommendations for any user's current web browsing sequence. A threshold called "minlen" is predefined which determines the minimum number of pages the user must visit prior to receiving recommendations from the website. Thus, only those users who visit more than "minlen" number of pages will be provided with recommendations. In the proposed system, the value of "minlen" is set to 2. When "minlen" numbers of pages are visited by the user, the browsing sequence of that particular user is passed as input to the recommendation engine. The recommendation engine then matches the current user access sequence with all the closed patterns in the closed pattern database until an appropriate match is found. There can be more than one pattern in the pattern database that matches with the current user web access sequence. When matching patterns are found, the recommendation engine recommends the next immediate pages following the current user web access sequence as recommendations to the user.

Recommendations Matching User Web Access Sequence

This block represents the navigation recommendations provided to the user by the recommendation engine. This block displays all the pages recommended by the navigation recommendation system. The user can select and visit any one web page of interest from amongst the several recommendedpages.

SOLUTION TO COLD START PROBLEM

The “Cold-Start” problem happens in recommendation systems due to lack of information, on new users or items. When a user is a new comer in a website he/she has not yet visited enough number of web pages. So, there is not enough evidence for the recommendation system to build the user profile based on his/her interests and the user profile will not be comparable to other users or items. As a result, the recommendation system cannot recommend any items to such a user.

However, the proposed system provides recommendations by matching the user’s current access sequence with the closed sequential patterns in the database and not by building the user profiles based on the interests of the users. Hence, in the proposed system navigation recommendations are provided even for new users.

Regarding the cold-start problem for web pages, when a web page is newly added to the website, users have not visited that page previously and hence it will not appear in the recommendation list of the website.

In the proposed system, new web pages are indicated by the “NEW” symbol so that the users know that the web page has been newly added and those interested can visit it. In this way, when the newly added web page is visited by several users it will ultimately get added in the recommendation list of the system.In this way, the cold start problem can be solved.

EVALUATION PARAMETERS FOR RECOMMENDATION SYSTEM

All the web pages recommended by the recommendation system as well as the recommended web pages visited by the user are stored in the database tables to evaluate the quality of the recommendations provided by the system.

The proposed system performs offline evaluation on the available data set to evaluate whether the recommendations provided by the system satisfy the expectations of the users. The proposed system is evaluated based on four parameters namely Precision, Recall, Fallout and Miss Rate as explained below.

Precision

Precision or true positive accuracy is calculated as the ratio of recommended items that are relevant to the total number of recommended items [6]. Precision is given as follows:

$$\text{Precision} = \frac{\text{Number of recommended pages visited}}{\text{Total number of recommended pages}}$$

This is the probability that a recommended item corresponds to the user’s preferences.

Recall

Recall or true positive is calculated as the ratio of recommended items that are relevant to the total number of relevant items [6]. Recall is given as follows:

$$\text{Recall} = \frac{\text{Number of recommended pages visited}}{\text{Total number of pages visited}}$$

This is the probability that a relevant web page is recommended.

Fallout

Fallout or false positive rate is calculated as the ratio of recommended items that are irrelevant to the total number of irrelevant items [6]. Fallout is given as follows:

$$\text{Fallout} = \frac{\text{Number of recommended pages not visited}}{\text{Total number of pages not visited}}$$

This is the probability that an irrelevant web page is recommended.

Miss Rate

Miss Rate or false negative rate is calculated as the ratio of items not recommended but actually relevant to the total number of relevant items [6]. Miss Rate is given as follows:

$$\text{Miss Rate} = \frac{\text{Number of pages not recommended but visited}}{\text{Total number of pages visited}}$$

This is the probability that a relevant web page is not recommended.

RESULTS AND DISCUSSIONS

The navigation recommendations provided by the proposed system are evaluated for accuracy of recommendations based on four parameters namely Precision, Recall, Fallout and Miss Rate. Table 1 shows the evaluation results for the different numbers of transactions.

Table 1: Evaluation Results of Proposed System

Number of Sessions	Precision	Recall	Fallout	Miss Rate
10	0.60	0.41	0.01	0.58
20	0.56	0.47	0.02	0.52
30	0.57	0.54	0.01	0.45
40	0.57	0.60	0.01	0.39
50	0.54	0.63	0.01	0.36

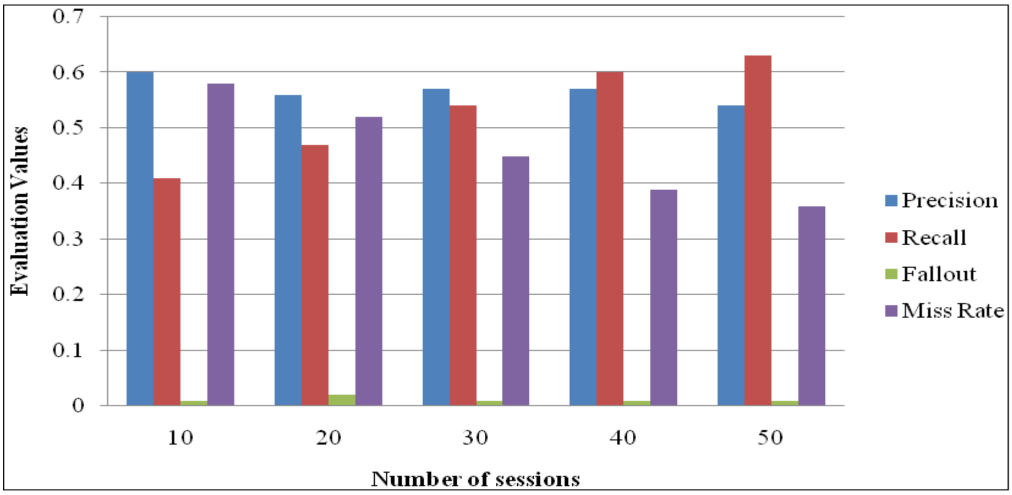

Figure 2: Evaluation Results of Proposed System

Table 1shows the values evaluated for Precision, Recall, Fallout and Miss Rate for different number of transactions. The value of Precision is highest i.e. 0.6 when the number of transactions is 10 and least i.e. 0.54 when the number of transactions is 50. The value of Recall is highest i.e. 0.63 when the number of transactions is 50 and least i.e. 0.41 when the number of transactions is 10. The value of Fallout is almost constant i.e. 0.01 for all set of transactions. The value of Miss Rate is highest i.e. 0.58 when the number of transactions is 10 and least i.e. 0.36 when the number of transactions is 50.

It is observed that the value of Recall increases as the number of transactions increases. At the same time, the value of Precision and Miss Rate decreases as the number of transactions increases. The value of Fallout is almost constant for any number of transactions.

CONCLUSIONS

The proposed recommendation system was implemented by integrating Preprocessing, Rough Set Clustering using upper approximation, Sequential Pattern Mining using Prefix Span algorithm and Closed Sequential Pattern Mining using Post-Pruning strategy.In the existing system, recommendations were generated using Sequential Pattern Mining using Prefix Span and Closed Sequential Pattern Mining using Post-Pruning strategy without clustering the user transactions. In the proposed system, Rough Set Clustering is performed prior to sequential pattern mining to improve the mining efficiency as well as to provide much accurate recommendations to users since clustering forms groups of all transactions having similar browsing patterns. The large number of patterns generated by Prefix Span, are then reduced using Closed Sequential Pattern Mining so that it becomes easy to handle the number of sequential patterns especially when the database to be mined is very large. When every user visits “minlen” number of web pages their current web access sequence is matched with the patterns in the database. All the patterns matching the current access sequence are provided as recommendations to the user.

On the basis of the evaluation results, it can be concluded that on an average about 57% of the recommended web pages correspond to the user preferences. Around 53% of the recommended web pages are relevant. About 1% of the recommended web pages are irrelevant and 46% of the relevant web pages are not recommended by the system.

Thus, the proposed recommendation system that performs clustering followed by Sequential Pattern Mining provides more efficient, accurate and effective recommendations to the users as compared to those that would be generated without performing clustering.

REFERENCES

1. Mitchell D’silva and Deepali Vora, “A Novel Recommendation System using Rough Set Clustering and Closed Sequential Pattern Mining”, International Journal of Computer Science Engineering and Information Technology Research, Vol. 3, Issue No. 3, August 2013, pp. 209 – 216.
2. Priyanka Patil and Ujwala Patil, “Preprocessing of Web Server log file for Web Mining”, World Journal of Science and Technology, In the Proceedings of “National Conference on Emerging Trends in Computer Technology, 21st April, 2012, pp. 14-18.
3. Ms. Jyoti, Dr. A. K. Sharma and Dr. Amit Goel. (2009).A novel Approach for Clustering Web User Sessions using RST. International Journal on Computer Science and Engineering (0975 - 3397), Vol. 2(1), pp. 56 – 61.
4. Anitha and Dr. N. Krishnan. (January2011).A Dynamic Web Mining Framework for E-learning

Recommendations using Rough Sets and Association Rule Mining, International Journal of Computer Applications (0975 - 8887),Vol. 12 – No.11,pp. 36 – 41.

5. Utpala Niranjan, Dr. R.B.V. Subramanyam and Dr. V. Khanaa. (May 2010). An Efficient System Based on Closed Sequential Patterns for Web Recommendations. International Journal of Computer Science Issues (1694 - 0784), Vol. 7, Issue 3, No. 4

6. Gunnar Schroder, Maik Thiele and Wolfgang Lehner, "Settings Goals and Choosing Metrics for Recommender System Evaluations". http://ucersti.ieis.tue.nl/files/papers/4.pdf